Searcha Co White Paper (August-2025)

Searcha Co / Ryan Pearce

2025-08-12

Searcha Co brings *indie grade* privacy and personalisation to web search. This non-technical white paper explains our mission, current architecture, and near-term roadmap for journalists, partners, and early adopters.

---

1 · Executive Summary

A one-person venture has indexed **2 billion** public-web documents and is preparing two complementary search engines:

- **Seek Ninja™** - fully stateless, ultra-private.
- **Searcha Page** - session-aware relevance that forgets you the moment the tab closes.

Both engines run on the same home-grown crawl → index → rank pipeline. All backend code is written and operated by a single developer; no venture funding, no reseller APIs.

---

2 · Indie Index at Web Scale

- **Corpus** 2 billion filtered pages, sourced primarily from recent Common Crawl dumps plus focused recrawls of high-change sites.
- **Storage** On server storage for performance. Image caching will begin in late 2025 with an opt-out removal process.
- **Hardware**-
  - *Primary node* - AMD EPYC 2nd gen | 512 GB RAM | 40 TB NVMe
  - *Support node* - desktop class CPU for auxiliary tasks
  - GPU work is rented on-demand; Searcha never self-hosts proprietary accelerators to keep pace with rapid LLM cycles.
- **CapEx** < US $4000 to date.
- **OpEx**-usage-based cloud GPUs + storage egress.

Searcha demonstrates that modern tooling can match early-2000s incumbents at a fraction of their budget - Google's celebrated "world's largest index" in 2000 held 1 billion URLs.[1]

[1] Google Press Center. "Google Launches World's Largest Search Engine" (June 26 2000).

3 · Privacy Without Trade-offs

Choice	Data stored	Lifetime	Intended user
Seek Ninja™	None - no cookies, no IDs, minimal server logs	Stateless	Maximum anonymity
Searcha Page™	One first-party session cookie (random ID)	Until tab or browser session ends (user-controlled)	Users who want short-term personalisation

Search queries are analysed in aggregate to improve ranking quality; no long term behavioural profiling is possible; IPs in logs are converted to a 1-way hash and resolution is reduced to a high collision rate. Both engines comply with GDPR and the California Consumer Privacy Act. Data removal requests can be sent to **chief@searcha.co**.

4 · Infrastructure Snapshot (August 2025)

Metric	Current	Target (post-upgrade)
Corpus size	2-B pages	3 B pages (Sep 2025)
P95 search latency (to first result)	**1.5-s**	1.0 s (after EPYC upgrade & prompt tuning)

Note on Latency: The recent reduction in latency from 2.5s to 1.5s was achieved through a significant architectural refactor, replacing a `POST/POLL` data retrieval pattern with **Server-Sent Events (SSE)**. This allows for a persistent connection and just-in-time data streaming to the client, greatly enhancing user experience. Further improvements will target the core search pipeline.

5 · Public-Beta Roadmap

Milestone	Target Date
Roll-out in-house “key-site” recrawler	**September 2025**
Index 3 B pages	**October 2025**
Third product (TBA)	**November 2025**
P95 latency 1 s	**December 2025**
Map & POI vertical	**Mid 2026**

6 · Compliance & Security

- **GDPR / CCPA** Data subjects can request removal of cached content via *chief@searcha.co*.
- **Security contact** `.well-known/security.txt` lists the same address for coordinated disclosure.
- **Image cache removal** Cache obeys `Cache-Control: no-store` and honours takedown requests within 48 h.

---

7 · Contact

Press & partnerships: **chief@searcha.co** (responses within 24 h). More info: https://searcha.co/

---

8 · Citations
